DOCUMENT RESUME

ED 379 342 TM 022 728

AUTHOR Reckase, Mark D.
TITLE Standard Setting on Performance Assessments: A Comparison between the Paper Selection Method and the Contrasting Groups Method.
PUB DATE Jun 94
NOTE 56p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Criteria; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Judges; Scores; Scoring; *Selection; *Statistical Distributions
IDENTIFIERS Contrasting Groups Method; National Assessment of Educational Progress; Paper Selection Method; *Performance Based Evaluation; Probabilistic Models; *Standard Setting

ABSTRACT

Comparative results are presented for procedures recently appearing in literature related to standard setting on the National Assessment of Educational Progress--the paper selection method and the contrasting group method. For this comparison, a probabilistic model with normal distribution of performance and a six-point scale were assumed. The paper selection method (American College Testing) requires judges to conceptualize students that are just at the borderline between categories. Papers are selected from a set that represents all levels of performance students at tne borderline would likely have produced. The standard is set at the mean of the scores assigned to those papers as part of the regular scoring process. The contrasting groups method requires that teachers internalize the construct to be assessed, and then select students above and below the criterion of success. Papers for those students are scored, and a point between the two score distributions is selected as the standard. For the cases that were modeled, the paper-selection procedure provided better estimates of the standard and the percent above the standard than did the contrasting group procedure. Six tables and 13 figures (graphs) present details of the analyses. (Contains 9 references.) (SLD)

Standard Setting on Performance Assessments: A Comparison between the Paper Selection Method and the Contrasting Groups Method[1]

Mark D. Reckase
ACT

Performance assessments are becoming more and more common as vehicles for determining the skills and knowledge acquired by students within the educational system within the United States. Aschbacher (1991) identified 13 states that had performance assessment procedures that were in place or in development, and another 10 states that were investigating the use of performance assessments. In the three years since that article, there have certainly been more state testing programs that have added performance assessments.

In addition to the uses in the states, the National Assessment of Educational Progress has embraced the use of performance assessment for its tests (e.g., see NAGB, 1992 for the framework of the reading test) and high stakes tests such as the SAT have added performance assessment components (Feryok & Wright, 1993). With the increased use of performance assessments for providing information in support of important judgements and decisions, there has been increased interest in setting decision points and standards on performance assessments. For example, the National Assessment Governing Board has defined target achievement levels as a means of quantifying the performance of the student population in the United States

[1]Paper presented at the National Conference on Large Scale Assessment sponsored by the Council of Chief State School Officers, Albuquerque, NM, June 1994.

(NAGB, 1990) and the April, 1994 meeting of the National Council on Measurement in Education included a symposium entitled "Setting Performance Standards on Complex Performance Assessments: Three Methods, Preliminary Results, and an Analytic Critique."

With the exception of the symposium mentioned above, there is very little in the measurement literature that provides guidance for setting standards on performance assessment tasks. Although some of the procedures described by Jaeger (1989) can be applied to tests composed of performance assessments, most of his chapter in *Educational Measurement, 3rd Edition* focuses on multiple-choice items or on total test scores without regard to the type of exercises used to produce the score. The purpose of this paper is to provide some initial insights into the challenges posed by setting standards on performance assessment tasks and to provide some comparative results for two procedures that have recently appeared in the literature related to standard setting on NAEP -- the paper-selection method (Luecht, 1993), and the contrasting group method (NAE, 1993). First, however, a brief description will be provided for the modeling process that was used to analyze the standard setting process and methods.

The Modeling Methodology

Since this is a preliminary investigation of the factors that affect the accuracy of standards set on performance assessments, it was considered important that the correct answer be known so that the results could be checked against a known

quantity. This belief limited the types of methodology that could be used for the investigation. Obviously, a real-data study could not be used, because, unfortunately, the true standard is unknown for such studies. The other alternatives, a simulation study, or a direct analysis of the characteristics of the standard setting process, were considered. The chosen methodology has some of the components of both simulations and direct analysis. A probabilistic model is proposed, but rather than use it to generate simulated data, the probabilistic model was analyzed directly to determine the implications for standard setting. Of course, the generalizability of the results will be dependent on the reasonableness of the model assumptions. Where ever possible, these assumptions were based on the real data from widely used testing programs.

The initial assumption of the model was that the performance of the examinee population was normally distributed and the performance task presented to the examinees was scored on a six-point scale. The six-point scale was selected because this type of scale is used to score many of the performance assessment tasks at ACT and was used for the NAEP Writing Assessment. For the analysis presented here, the performance on the task for the student population was assumed to have a mean of 3.5 and a standard deviation of 1.0. The continuous normal distribution was converted to a discrete distribution for all further analyses (see Figure 1).

Insert Figure 1 about here

This distribution has been labeled the "capability" distribution because it is considered to be the true score distribution for students that are making a reasonable effort to perform well. However, students will not always perform exactly as the capability distribution would indicate. Sometimes, they might not apply their full capabilities to the task and produce work that is poorer in quality than they could produce under different circumstances. At other times, they may expend extraordinary effort and outperform their usual level of quality.

To model the students' actual performance, rather than their capability, a probability distribution was defined for each score level in the original distribution. These distributions show the likelihood that a student at capability, c, will perform at each capability level. Since these are conditional probabilities, they will be represented symbolically as $P_{j|i}$, indicating that the student with capability i has probability $P_{j|i}$ of producing a paper at level j. Table 1 provides the performance distributions for students at each capability level.

Insert Table 1 about here

The model assumed here specifies that 80% of the students will perform at their level of capability. However, approximately 15% score lower than their level of capability and 5% score at a higher level. Of course, very low stakes tests might have more extreme spread in the distributions and very high stakes tests could have more students performing beyond their normal capabilities.

The student performance distributions and the capability distribution can be combined using standard probability theory to derive the expected performance distribution for the total student population. That is, the proportion of students with performance in score category i, accumulated over all performance distributions is given by

$$P(s=i) = \sum_{j=1}^{6} P(c=j)\, P_{i|j}$$

where P(s=i) is the proportion of students with performance level i,

P(c=j) is the proportion of students with capability j,

and $P_{i|j}$ is the probability that a student with capability j will perform at level i.

The resulting performance distribution is given in Table 4 and in Figure 2. Since most students are expected to perform up to their capabilities, the performance distribution is not very different from the capability distribution, but it is shifted slightly toward the lower end of the scale. If the drop in performance is expected to be greater, the

probabilities in Table 1 can be adjusted to reflect that drop and a new performance distribution can be obtained.

The distribution given in Figure 2 is the distribution of scores that would be obtained if the papers the students produced were scored without error. However, from long experience in the scoring of essays, it is clear that there is substantial error in scoring, even with well trained scorers. Dunbar, Koretz & Hoover (1991) summarized some of the information that is available on scoring reliability. Given those results, it seems reasonable to assume that the reliability of scoring might be about .5. That level of reliability results in a standard error of measurement of .724 if the score variance is 1.0.

To generate the expected observed score distribution for the scored student performance, a possible model is to assume that the scorers are statistically unbiased with a standard error of .724. Those assumptions were used to generate the probability that an observed score would be assigned, given the true performance level. These probabilities are presented in Table 2. At the extremes, when assigned scores were below 1 or above 6, they were placed in the 1 or 6 category respectively.

Insert Table 2 about here

Given the probabilities in Table 2 and the performance distribution in Figure 2, an expected observed score distribution can be computed using the direct analogue to

the equation provided above. The resulting score distribution is presented in Figure 3. Clearly, the distribution has become flatter with heavier tails. This is exactly the results that would be predicted from classical test theory.

Insert Figure 3 about here

To provide some context for this model of performance and scoring of performance assessments, the observed score distribution for one of the NAEP Writing Samples is presented in Figure 4. This particular example was selected because it had the best combination of spread of scores and mean score near 3.5. Note that it is quite a bit more peaked and less spread that the observed score distribution given in Figure 3. For easier comparison of the distributions, the proportion in each score category, the means, and standard deviations are presented in Table 4.

Insert Figure 4 and Table 4 about here

Since the intent of the process that has just been described is to accurately model observed performance assessment score distributions, the potential reasons for the differences in Figures 3 and 4 are of interest. Three possible reasons seem possible. First, the scoring of the NAEP Writing Sample might be more reliable than

was assumed here. However, the true score distributions shown in Figures 1 and 2 are less peaked and have heavier tails, so merely reducing the standard error due to scoring will not result in a model distribution that matches the NAEP distribution.

A second possible explanation for the differences in the distributions is that the true score distribution has a smaller standard deviation. Some preliminary work showed that reproducing Figure 4 would require almost all true scores to be in categories 3 and 4 if all other assumptions were maintained. While it is possible that there are only two categories of papers, that seemed like an unlikely occurrence, so a third possibility was considered.

The third explanation is that the rating of the papers suffered from a regression effect caused by a reluctance on the part of the readers to give extreme scores. To model a scorer regression effect, the true scores were regressed to the mean by multiplying the distance from the mean by .75 and then the error distributions were placed around the regressed score. The resulting probabilities of observed scores for a given true performance level are presented in Table 3.

Insert Table 3 about here

The observed score distribution that results from the combination of regression and a standard error of .724 is presented in Figure 5 and in the last row of Table 4. This distribution is closer to that of the observed NAEP distribution, but it still does not

reproduce it exactly. It seems that the NAEP Writing Sample distribution has a stronger regression effect for high scores than for low ones, or that performance is lower overall. However, for the purposes of this paper, the distribution shown in Figure 5 will be considered as a reasonable approximation to the results that might be obtained on an actual performance assessment task.

Standard Setting

The description of the modeling methodology is but a means to set the stage for the focus of this paper, the setting of standards on performance assessment items. Suppose that a testing program contains test tasks that are scored holistically using a six-point rubric and that it is necessary to set performance standards on these test tasks. Two approaches to the setting of standards have been proposed in recent literature on NAEP. The first is the "paper selection method" (American College Testing, 1993). This method requires judges to conceptualize students that are just at the borderline between categories. They are then to select papers from a set that represents all levels of performance that students at the borderline would likely have produced. The standard is set at the mean of the scores assigned to those papers as part of the regular scoring process.

The second method is the contrasting groups method (National Academy of Education, 1993). This method requires that teachers first internalize the construct to be assessed and then select students that are above and below the criterion of success. The papers for those students are scored and a point between the two

score distributions is selected as the standard. The question posed in this paper is which of these two methods more accurately reproduces a known, true standard when all sources of error are accounted for in the performance assessment process?

Paper Selection Method

To model the paper selection method, it was first necessary to make assumptions about how well judges could select papers. It seemed reasonable that standard-setting judges, even with extensive training, would not be able to select papers more accurately than trained scorers. Therefore, it was assumed that judges would select papers with a standard error of .724, the same value used to model the error in scoring papers. For all cases, the true standard will be assumed to be 4.0 on the six-point scale. This standard results in 31% of the original population above the standard and 69% below the standard.

Given the assumed accuracy of the judges paper selections, the true score distribution for the papers selected by the judges is given in Figure 6. The difference between student capabilities and performance is not of interest here because the judges only look at performance. However, estimating the percent that is above the final standard can only be done from the observed score distribution that includes all the sources of error.

Insert Figure 6 about here

The judges would select the papers without seeing the scores provided by the usual scoring process. After the papers are selected, the observed score distributions can be produced. Assuming no regression effect and the same scorer error as above, the expected observed score distribution is given in Figure 7. This distribution is fairly symmetric around the score of 4.0 and has a mean of 3.99. Thus, if there is no regression of scores to the mean, the paper selection method yields very accurate results.

Insert Figure 7 about here

Figure 8 presents the expected observed score distribution when scorer regression to the mean is present. The mean of that distribution is 3.87. Thus, the type of scorer error that contains a regression effect results in setting a slightly lower standard even when the judges are unbiased in the paper selections.

Contrasting Groups

Modeling the contrasting groups procedure is quite a bit more complicated. First, the true score distribution was divided at the value of 4.0 into two distributions: one for passing students, and the other for failing students. These two distributions are presented in Figure 9. Note that these two distributions both contain the 4 score point.

Insert Figure 9 about here

Next, it is unlikely that the teachers will classify the students into the two categories without error. Since the teachers will be very familiar with the students' work, a smaller standard error, .5, was assumed for the ratings. This standard error is equivalent to a scoring reliability of .75. The distributions for the passing and failing groups including classification error are given in Figure 10. These are still true score distributions. For the failing group, those with scores of 5 and some of those with 4 are misclassified. Likewise, for the passing group, those with scores of 3 and some of those with 4 are misclassified. The misclassification rate is about 10%.

Insert Figure 10 about here

When the students take the performance assessment, there may be a difference between their capabilities and their actual performance. Therefore the capability distributions were modified in the way described above to give the expected performance distributions. The distributions for the failing and passing groups are presented in Figure 11. Note that the overlap between the two distributions has increased.

Insert Figure 11 about here

Finally, the scoring error needed to be included in the distributions. Figure 12 gives the distributions with no scorer regression, but with the same scorer error used earlier, and Figure 13 gives the distributions with scorer error and the .75 regression effect. The Figures also give the means for the two groups for the regression and no regression conditions. Table 5 provides a summary of the full set of distributions.

Inset Figures 12 and 13 and Table 5 about here

There is no single method that is recommended for setting a standard once the distributions for the students classified as passing and failing have been determined. Several methods consider the point where the score distributions cross, or where the number of passing scores for the two groups in a given score category are equal (Jaeger, 1989). These method appear to be sensitive to the number of examinees in each group. In this case, there are about twice as many failing students as passing students. Different standards will be set using the above methods if the observed distributions are used as is, or if they are converted to relative frequency distributions so that the two distributions have the same overall area.

To overcome the problem of the difference in the relative sizes of the distributions, the procedure for setting the standard in this case was to pick the point that was midway between the means for the two distributions. The belief was that the same value would be estimated regardless of the number of examinees in each group. This process for setting the standard will yield roughly the same result as the intersection of the two distributions, if the distributions are scaled to have the same number of examinees in each.

The standards for the contrasting group method were computed for both the case with no scorer regression, and assuming the .75 scorer regression. These results are presented in Table 6, along with a summary of the results for the paper-selection method and the true standard, the value used to generate the distributions. Note that the standards based on the contrasting group methods are closer to the mean of the total population than the paper-selection method. They also underestimate the true value.

Inset Table 6 about here

In some sense, the relationship to the true standard is irrelevant because that value is never known in practice. A value that is of more interest is the percentage of examinees that will be above the standard. The true number of students above the standard based on capabilities is 31% and based upon performance is 29%. Which of

these two values is of greater importance depends on the purpose of the testing program. In either case, however, the number of passing students is about 30%. If the standards are set on data containing all of the sources of error, which will yield a standard that is closest to providing the correct percentage above the standard?

For the paper-selection procedure the percent above the standard is 32% if there is no regression bias in the scoring and 34% if there is regression in the scoring process. Thus, paper-selection slightly overestimates the percent passing for the situation modeled here. The contrasting group method, on the other hand provides estimates of 42% and 43%, respectively. This is quite a large overestimate of the proportion passing. If the cases modeled here are at all typical, these results suggest a strong statistical bias in the contrasting groups procedure.

While it is impossible to generalize too far beyond the particular situations modeled here, these results do suggest that the contrasting groups method should be used with caution until the factors that affect the standards are better understood. The results reported here are likely due, at least in part, to the statistical regression of the scores of the upper group toward the mean. This effect is likely to be present any time one of the groups is much farther from the mean than the other.

Conclusions

This paper presents a methodology for studying the characteristics of standard setting procedures. The intent of the method is to model the sources of error for the various methods. The results of this study are generalizable only to the extent that

the assumptions about error distributions are reasonable. However, many different types of error distributions can be tried and if the results are insensitive to the assumptions that are made, some confidence can be placed in the results.

The basic finding of the study was that, for the cases that were modeled, the paper-selection procedure provided better estimates of the standard and the percent above the standard than did the contrasting group procedure. The latter procedure underestimated the standard and overestimated the number of examinees above the standard.

Before considering these results as indicating a serious problem with the contrasting groups method, many more cases should be investigated, and the validity of the assumptions for actual performance assessment items should be tested. Until that is accomplished, these results provide food for thought.

References

American College Testing (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Writing: final report.* Iowa City, IA: Author.

Aschbacher, P. R. (1991). Performance assessment: state activity, interest, and concerns. *Applied Measurement in Education, 4(4),* 275-288.

Dunbar, S. B., Koretz, D. M. & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4(4),* 289-303.

Feryok, N. J. & Wright, N. K. (1993, April). Design of the SAT and PSAT/NMSQT field trial. Paper presented at the meeting of the National Council of Measurement in Education, Atlanta.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) *Educational Measurement (3rd Ed.)*. New York: American Council on Education and Macmillan.

Luecht, R. M. (1993, April). Using IRT to improve the standard setting process for dichotomous and polytomous items. Paper presented at the meeting of the National Council of Measurement in Education, Atlanta.

National Academy of Education (1993). *Setting performance standards for student achievement*. National Center for Educational Statistics, Washington, DC: U.S. Government Printing Office.

National Assessment Governing Board (1990). *NAGB achievement levels policy (Appendix A in Request for Proposals)*. Washington, DC: Author.

National Assessment Governing Board (1992). *Reading framework for the 1992 National Assessment of Educational Progress*. Washington, DC: Author.

Table 1
Performance Distribution for Each Capability Level

Capability Level	Performance Level					
	1	2	3	4	5	6
1	.96	.05	0	0	0	0
2	.15	.80	.05	0	0	0
3	.05	.10	.80	.05	0	0
4	0	.05	.10	.80	.05	0
5	0	0	.05	.10	.80	.05
6	0	0	0	.05	.15	.80

Table 3
Scoring Error Distribution
SEM = .724, Regressed .75 to Mean

True Performance Level	Observed Score					
	1	2	3	4	5	6
1	.43	.46	.11	0	0	0
2	.11	.46	.37	.06	0	0
3	.01	.18	.51	.27	.03	0
4	0	.03	.27	.51	.18	.01
5	0	0	.06	.37	.46	.18
6	0	0	0	.11	.46	.43

Table 2
Scoring Error Distribution
SEM = .724 r_{xx} = .5

True Performance Level	Reported Score					
	1	2	3	4	5	6
1	.75	.23	.02	0	0	0
2	.25	.50	.23	.02	0	0
3	.02	.23	.50	.23	.02	0
4	0	.02	.23	.50	.23	.02
5	0	0	.02	.23	.50	.25
6	0	0	0	.02	.23	.75

Table 4
Summary of Capability, Performance Observed, and a NAEP Distribution

Distribution	Score Category							
	1	2	3	4	5	6	$\bar{x}$	SD
Capability	.02	.14	.34	.34	.14	.02	3.5	1.025
Performance	.057	.164	.320	.304	.132	.023	3.36	1.141
Observed	.090	.175	.271	.260	.148	.056	3.37	1.325
NAEP	.015	.191	.434	.305	.052	.004	3.43	.951
Regressed	.046	.168	.320	.303	.136	.027	3.39	1.134

Table 5
Contrasting Groups
Score Distributions

Distribution Type	Failing						Passing					
	1	2	3	4	5	6	1	2	3	4	5	6
True	.02	.14	.34	.19	0	0	0	0	0	.15	.14	.02
Classification	.02	.14	.333	.178	.003	0	0	0	.007	.162	.137	.02
Performance	.051	.155	.291	.159	.011	0	0	.009	.029	.145	.121	.023
No Regression	.087	.161	.219	.152	.048	.006	.003	.014	.052	.108	.100	.05
Regression	.044	.155	.256	.173	.042	.003	.001	.014	.064	.129	.093	.025

Table 6
Standards and Percent Passing for Various Conditions

Case	Standard	Distribution			
		Capability	Performance	No Regression	Regression
True	4.0	31	29	32	29
Paper Selection					
No Regression	3.99	32	29	32	30
Regression	3.87	36	33	35	34
Contrasting Groups					
No Regression	3.62	45	41	42	42
Regression	3.58	47	42	44	43

Figure 1
True Capability Distribution

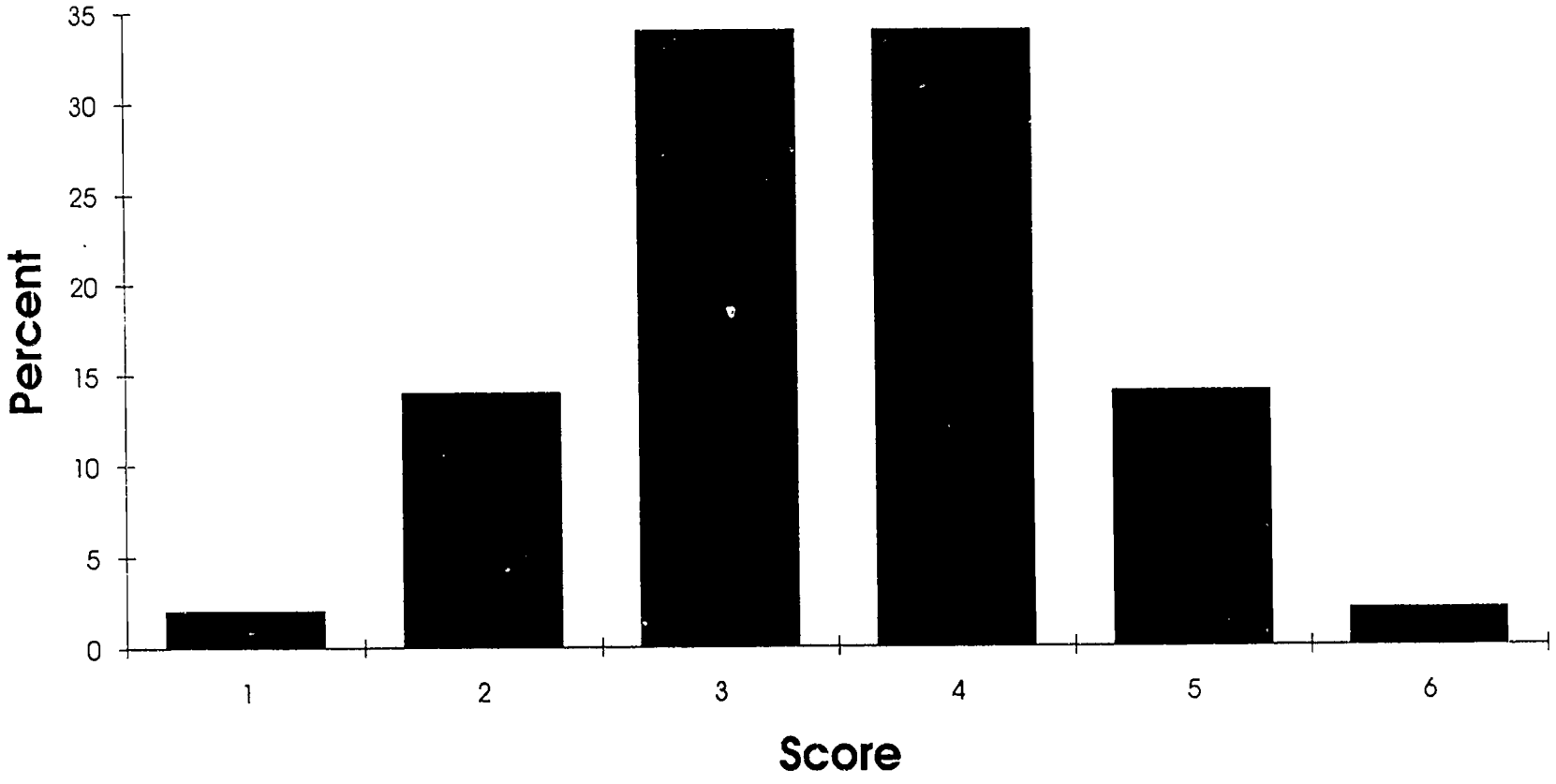

Figure 2
Performance Distribution

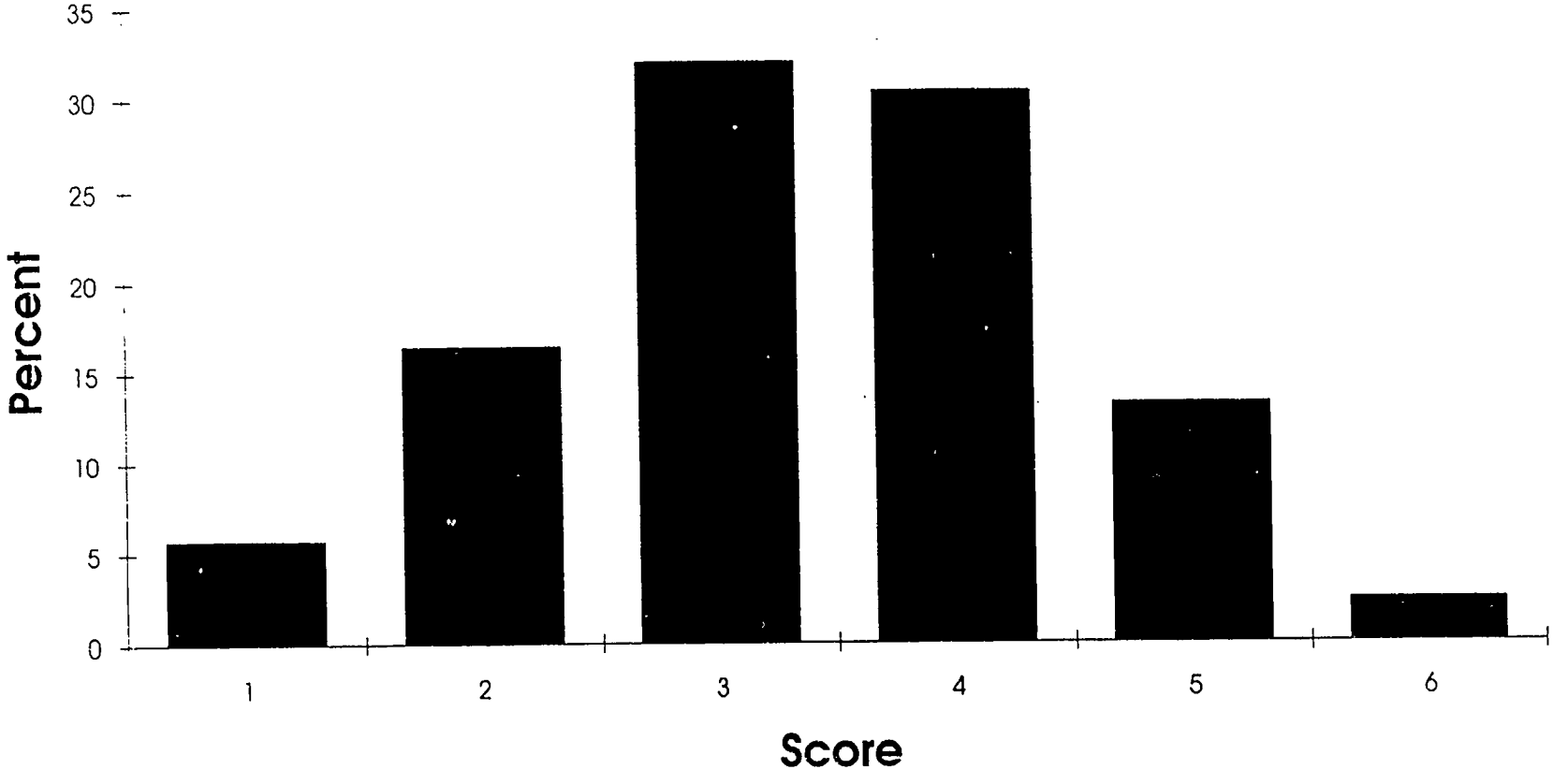

Figure 3
Observed Score Distribution

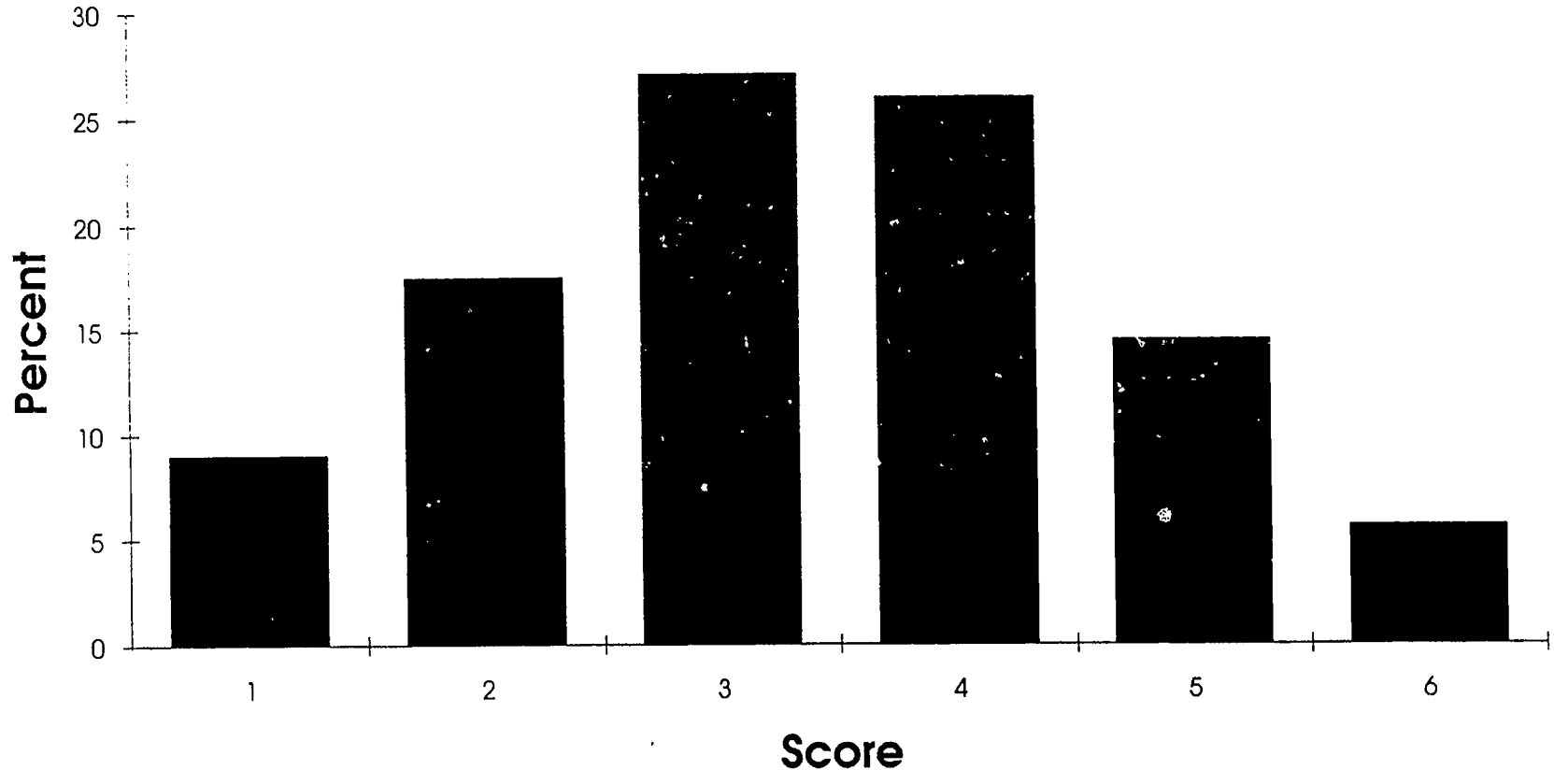

Figure 4
NAEP Writing Sample Distribution

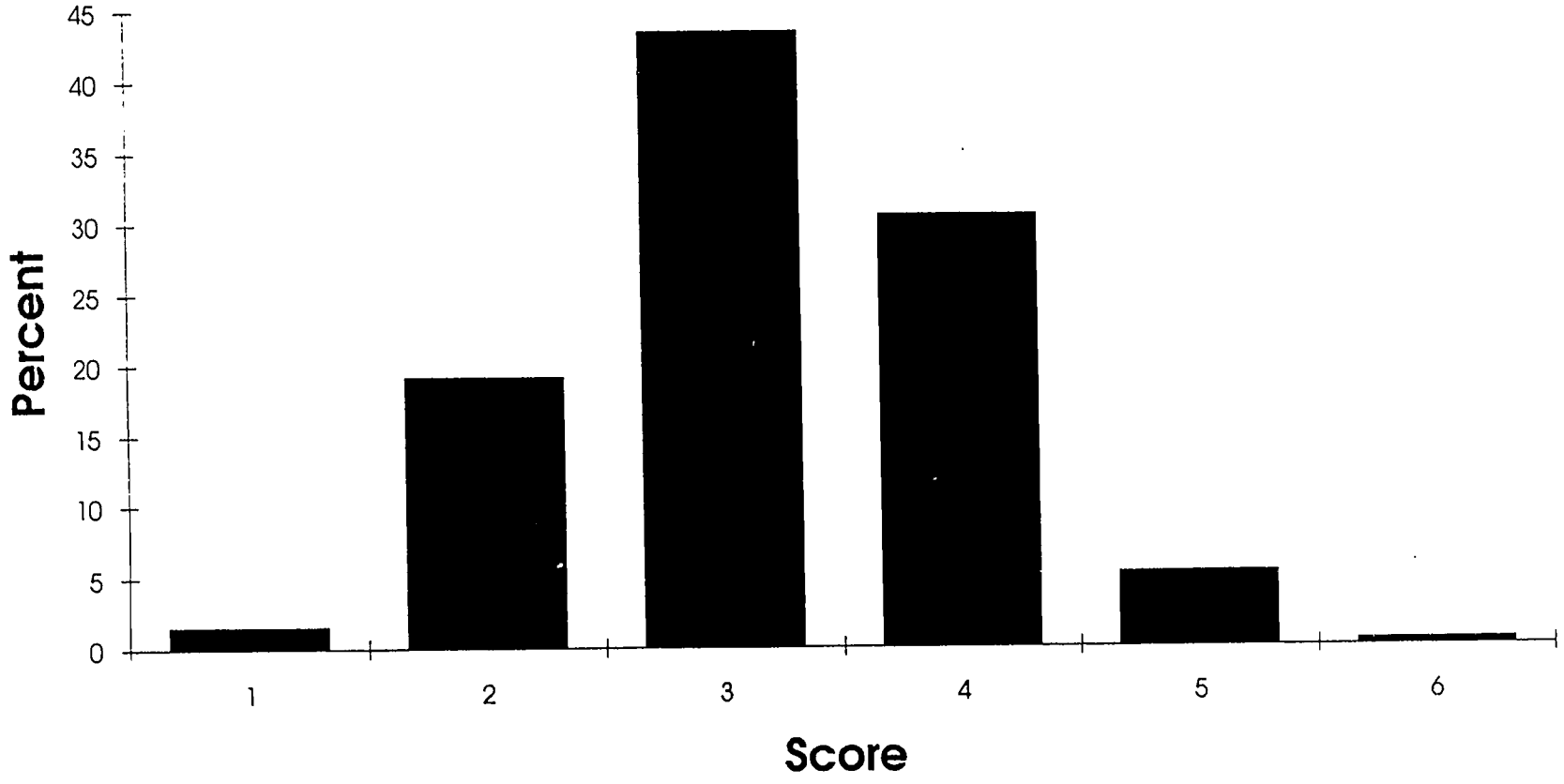

Figure 5
Observed Score Distribution
Scorer Regression

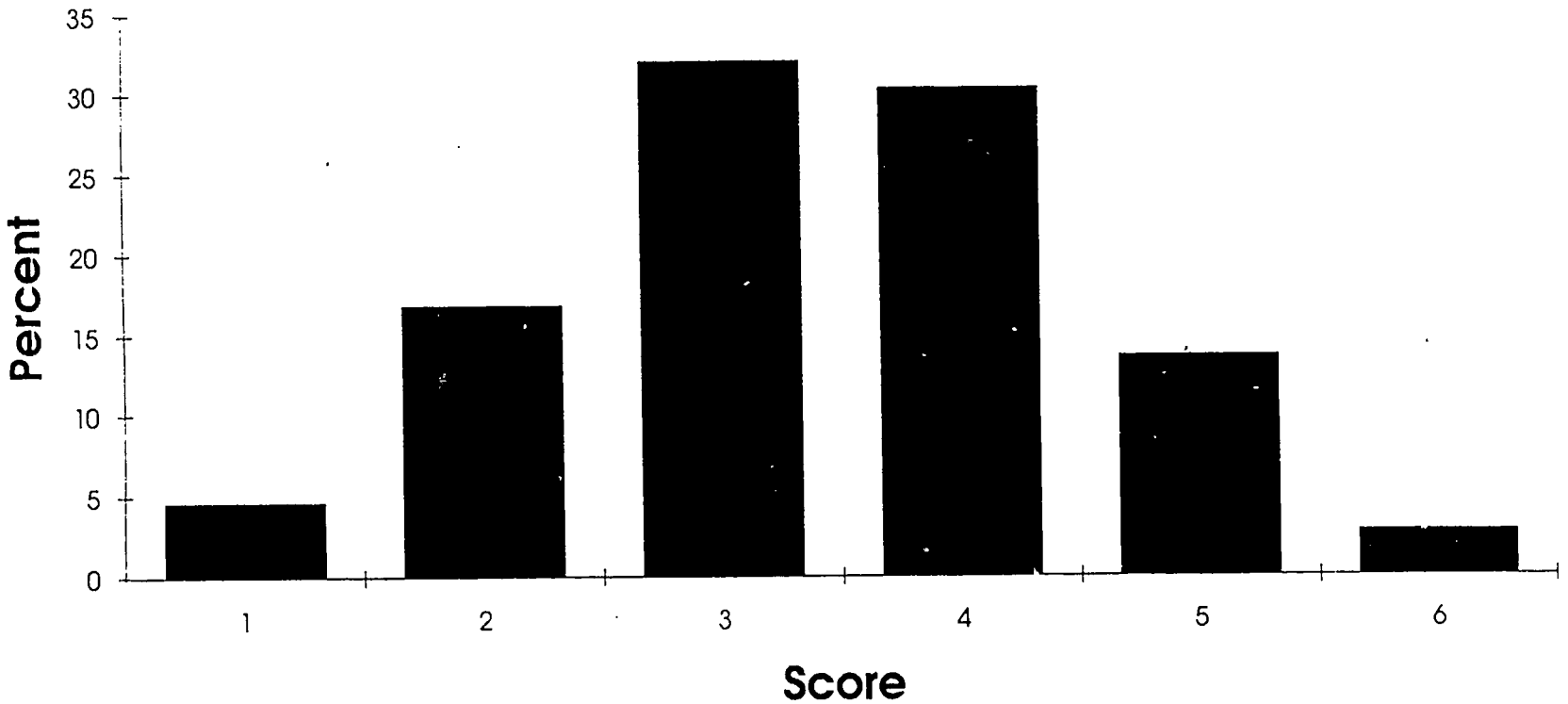

Figure 6
Distribution of Judges Paper Selections
SEM = .724 TRUE STANDARD 4.0

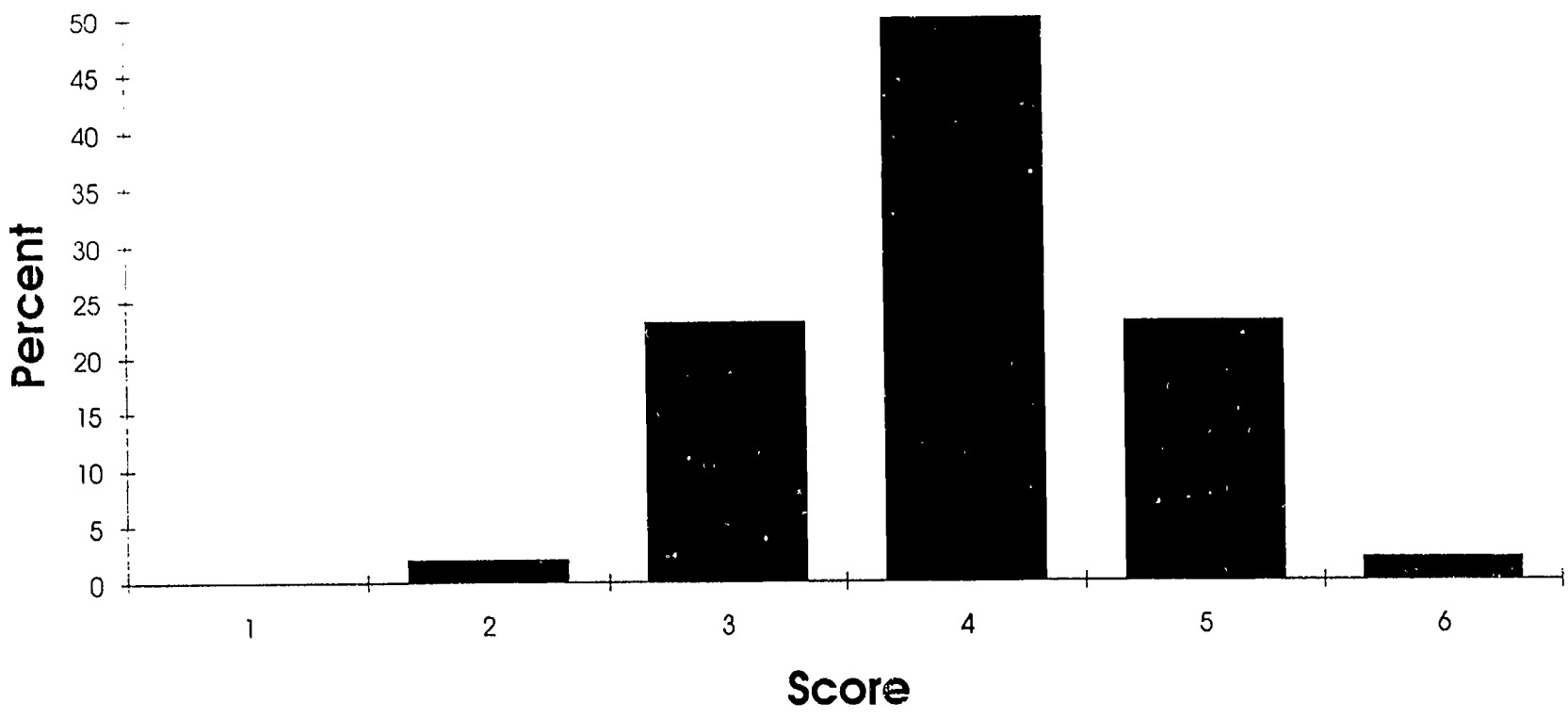

Figure 7
Observed Score Distribution
SEM = .724 No Regression

MEAN = 3.99

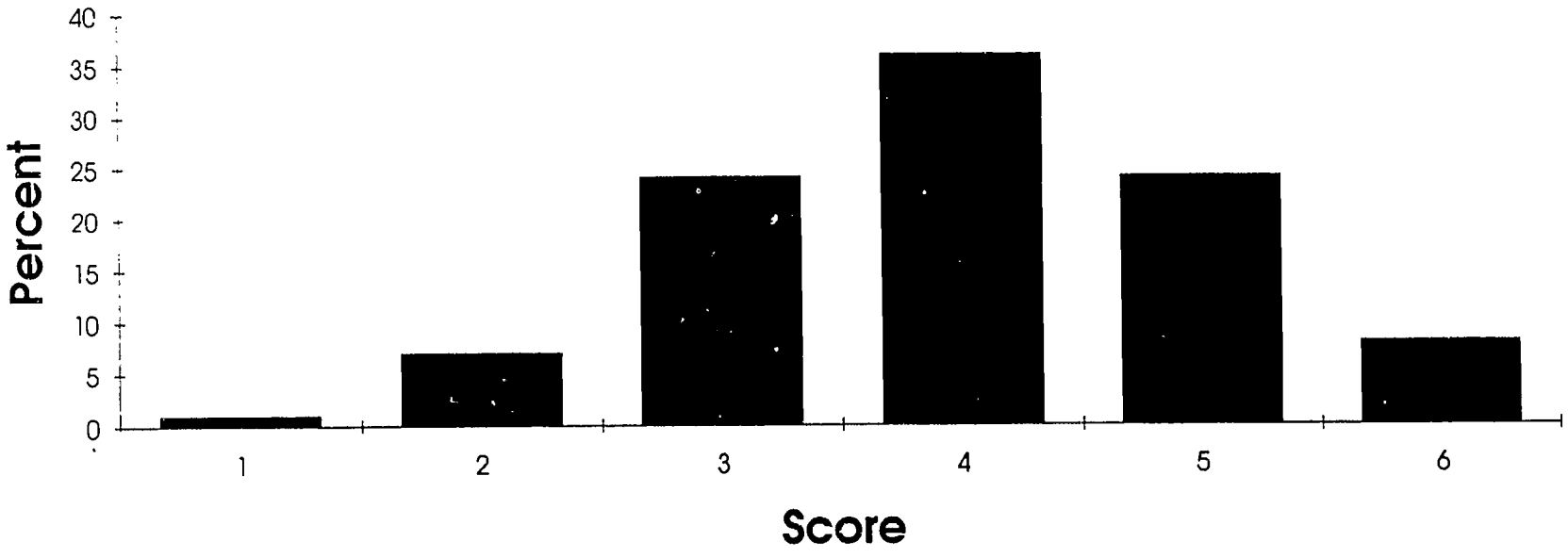

Figure 8
Observed Scores for Judgements
SEM = .724 with Scorer Regression

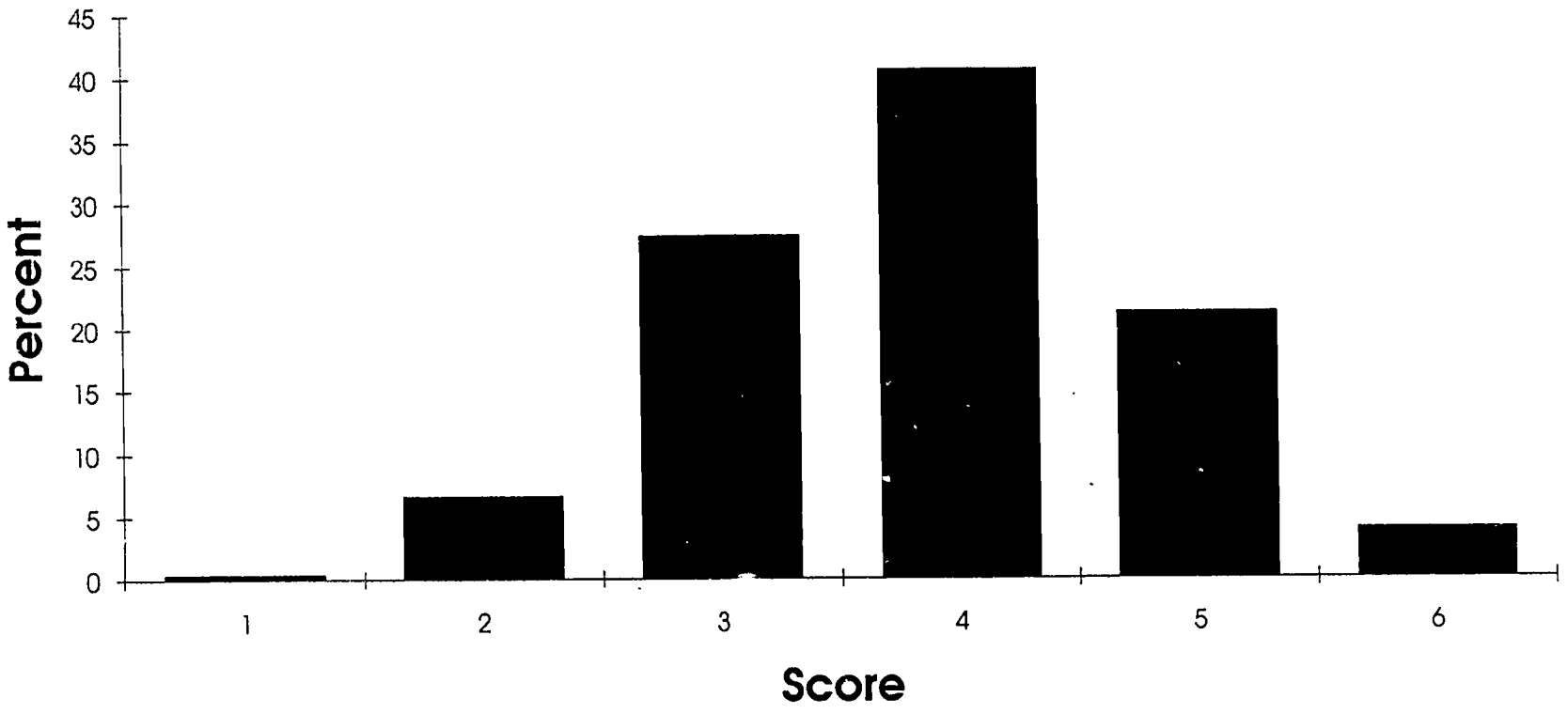

Figure 9
Contrasting Groups True Distribution
True Standard 4.0

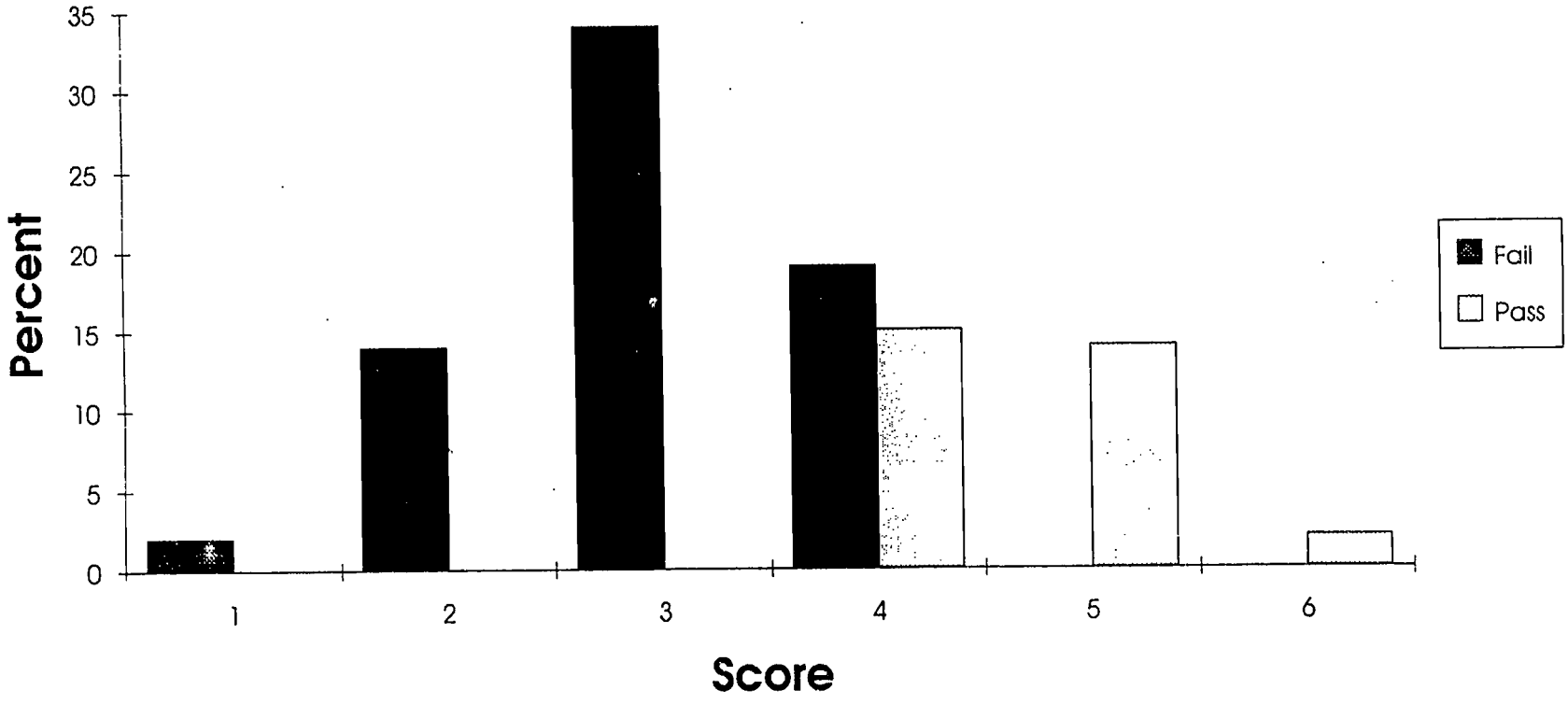

Figure 10
Contrasting Groups with Classification Error
Rater SEM = .5

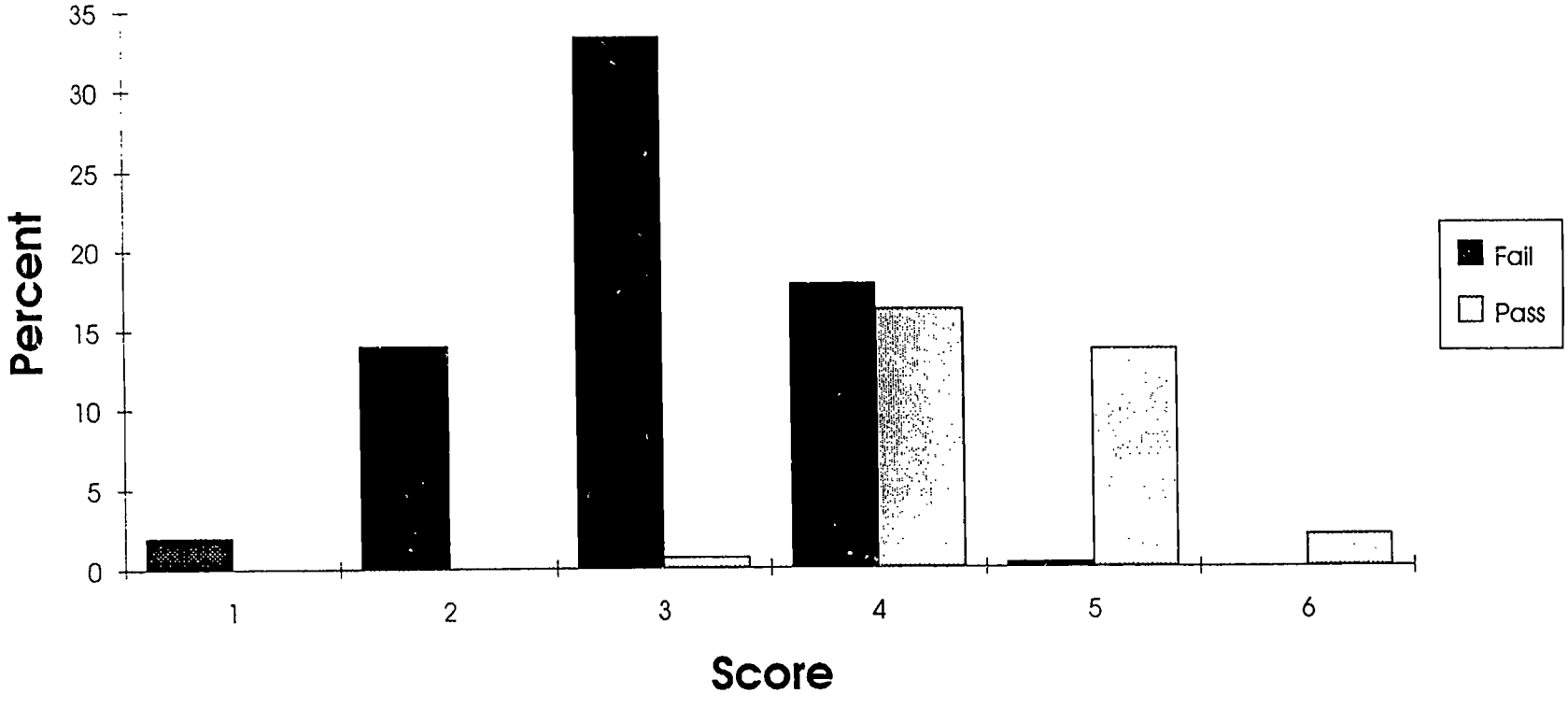

Figure 11
Contrasting Groups Performance Distribution

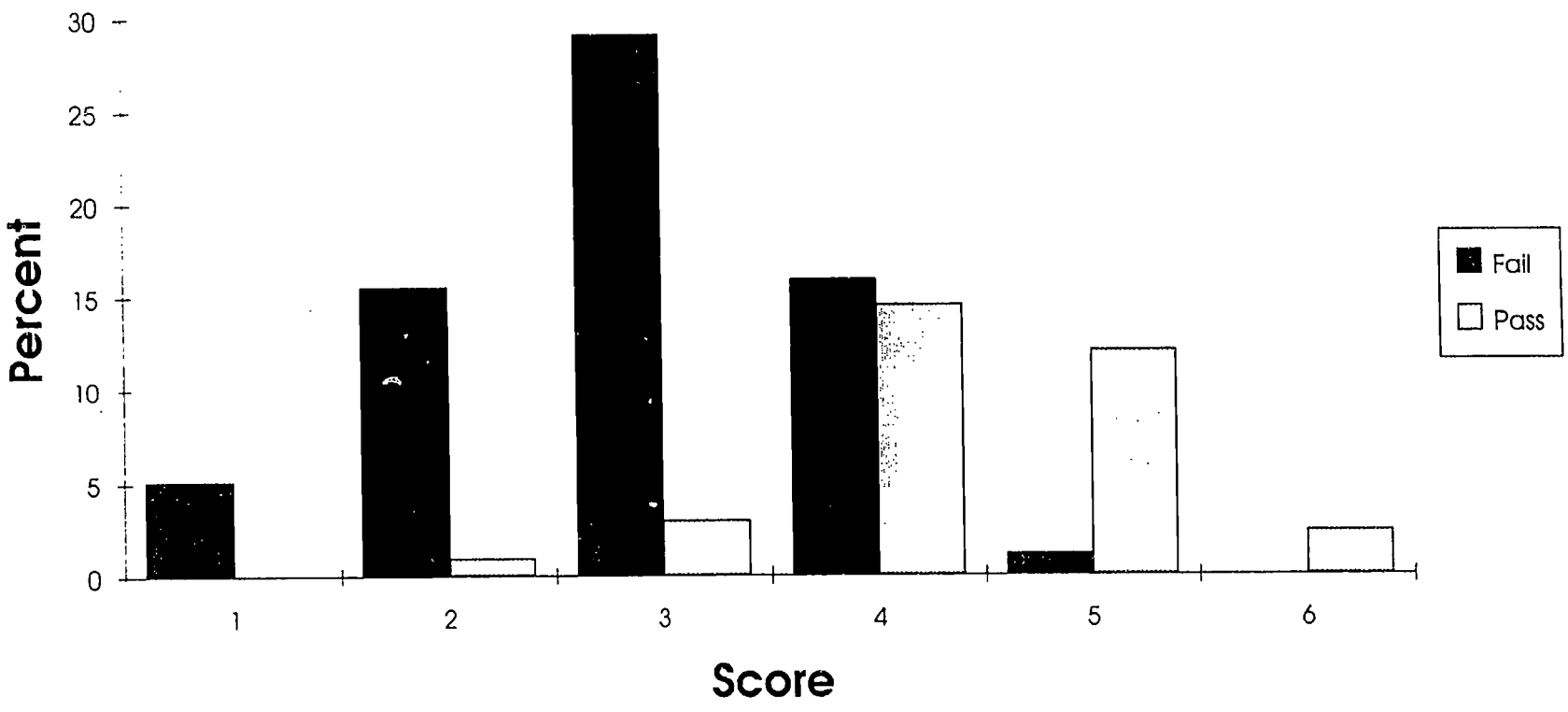

Figure 12
Contrasting Groups
Observed Score Distribution
SEM = .724 No Regression

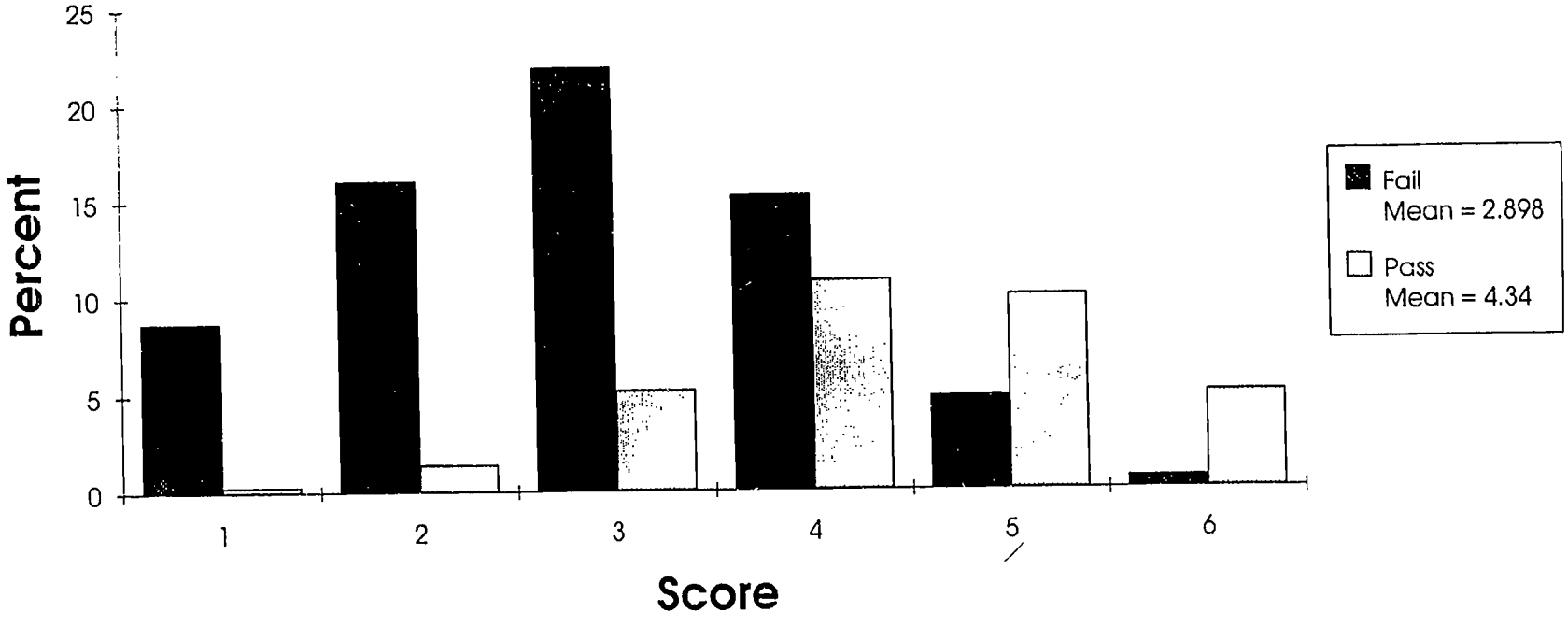

Figure 13
Contrasting Groups
Observed Score Distribution
SEM = .724 Scorer Regression